# Application of Information Theory, Lecture 2
# Joint & Conditional Entropy, Mutual Information

**Handout Mode**

Iftach Haitner

Tel Aviv University.

Oct 27, 2015

Part I

# Joint and Conditional Entropy

# Joint entropy

- Recall that the entropy of rv $X$ , is defined by

$$H(X) = -\sum_{x \in \mathcal{X}} \mathrm{P}_X(x) \log \mathrm{P}_X(x)$$

- Shorter notation: for $X \sim p$, let $H(X) = -\sum_x p(x) \log p(x)$ (where the summation is over the domain of $X$).
- The joint entropy of (jointly distributed) rvs $X$ and $Y$ with $(X, Y) \sim p$, is

$$H(X, Y) = -\sum_{x,y} p(x, y) \log p(x, y)$$

This is simply the entropy of the rv $Z = (X, Y)$.

- Example:

| $X \backslash Y$ | 0 | 1 |
|---|---|---|
| 0 | $\frac{1}{4}$ | $\frac{1}{4}$ |
| 1 | $\frac{1}{2}$ | 0 |

$$\begin{aligned} H(X, Y) &= -\frac{1}{2} \log \frac{1}{\frac{1}{2}} - \frac{1}{4} \log \frac{1}{\frac{1}{4}} - \frac{1}{4} \log \frac{1}{\frac{1}{4}} \\ &= \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 2 = 1\frac{1}{2} \end{aligned}$$

## Joint entropy, cont.

- The joint entropy of $(X_1, \ldots, X_n) \sim p$, is

$$H(X_1, \ldots, X_n) = -\sum_{x_1, \ldots, x_n} p(x_1, \ldots, x_n) \log p(x_1, \ldots, x_n)$$

# Conditional entropy

- Let $(X,Y) \sim p$, let $p_X = \sum_y p(x,y)$, $p_Y = \sum_x p(x,y)$ and $p_{Y|X}(y|x) = \frac{p(x,y)}{p_Y(y)}$.
- For $x \in \mathrm{Supp}(X)$, the random variable $Y|_{X=x}$ is well defined (distributed according to $q(y) = p_{Y|X}(y|x)$).
- The entropy of $Y$ conditioned on $X$, is defined by

$$H(Y|X) := \mathop{\mathrm{E}}_{x \leftarrow X} H(Y|_{X=x})$$

- Measures the uncertainty in $Y$ given $X$.
- $$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{X}} p_X(x) \cdot H(Y|_{X=x}) \\ &= -\sum_{x\in\mathcal{X}} p_X(x) \sum_{y\in\mathcal{Y}} p_{Y|X}(y|x) \log p_{Y|X}(y|x) \\ &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y) \log p_{Y|X}(y|x) \\ &= -\mathop{\mathrm{E}}_{(X,Y)} \log p_{Y|X}(Y|X) = -\mathop{\mathrm{E}}_{Z=p_{Y|X}(Y|X)} \log Z \end{aligned}$$

# Conditional entropy, cont.

- Example

| $X \backslash Y$ | 0 | 1 |
|---|---|---|
| 0 | $\frac{1}{4}$ | $\frac{1}{4}$ |
| 1 | $\frac{1}{2}$ | 0 |

What is $H(Y|X)$ and $H(X|Y)$?

$$\begin{aligned} H(Y|X) &= \mathop{\mathrm{E}}_{x \leftarrow X} H(Y|_{X=x}) \\ &= \frac{1}{2} H(Y|_{X=0}) + \frac{1}{2} H(Y|_{X=1}) \\ &= \frac{1}{2} H(\frac{1}{2}, \frac{1}{2}) + \frac{1}{2} H(1,0) = \frac{1}{2}. \end{aligned}$$

$$\begin{aligned} H(X|Y) &= \mathop{\mathrm{E}}_{y \leftarrow Y} H(X|_{Y=y}) \\ &= \frac{3}{4} H(X|_{Y=0}) + \frac{1}{4} H(X|_{Y=1}) \\ &= \frac{3}{4} H(\frac{1}{3}, \frac{2}{3}) + \frac{1}{4} H(1,0) = 0.6887 \neq H(Y|X). \end{aligned}$$

## Conditional entropy, cont..

- $$\begin{aligned} H(X|Y,Z) &= \mathop{\mathrm{E}}_{(y,z)\leftarrow(Y,Z)} H(X|_{Y=y,Z=z}) \\ &= \mathop{\mathrm{E}}_{y\leftarrow Y}\ \mathop{\mathrm{E}}_{z\leftarrow Z|_{Y=y}} H(X|_{Y=y,Z=z}) \\ &= \mathop{\mathrm{E}}_{y\leftarrow Y}\ \mathop{\mathrm{E}}_{z\leftarrow Z|_{Y=y}} H((X|_{Y=y})|_{Z=z}) \end{aligned}$$

  Let $(X_y, Z_y) = (X,Z)|_{Y=y}$. Then

  $$\begin{aligned} H(X|Y,Z) &= \mathop{\mathrm{E}}_{y\leftarrow Y}\ \mathop{\mathrm{E}}_{z\leftarrow Z_y} H(X_y|_{Z=z}) \\ &= \mathop{\mathrm{E}}_{y\leftarrow Y}\ \mathop{\mathrm{E}}_{z\leftarrow Z_y} H(X_y|_{Z_y=z}) \\ &= \mathop{\mathrm{E}}_{y\leftarrow Y} H(X_y|Z_y) \end{aligned}$$

## Relating mutual entropy to conditional entropy

- What is the relation between $H(X)$, $H(Y)$, $H(X,Y)$ and $H(Y|X)$?
- Intuitively, $0 \leq H(Y|X) \leq H(Y)$

  Non-negativity is immediate. We prove upperbound later.
- We will also see that $H(Y|X) = H(Y)$ iff $X$ and $Y$ are independent.
- In our example, $H(Y) = H(\frac{3}{4}, \frac{1}{4}) > \frac{1}{2} = H(Y|X)$
- Note that $H(Y|X = x)$ might be larger than $H(Y)$ for some $x \in \mathrm{Supp}(X)$.
- Chain rule (proved next). $H(X,Y) = H(X) + H(Y|X)$
- Intuitively, uncertainty in $(X,Y)$ is the uncertainty in $X$ plus the uncertainty in $Y$ given $X$.
- $H(Y|X) = H(X,Y) - H(X)$ is as an alternative definition for $H(Y|X)$.

## Chain rule (for the entropy function)

**Claim 1**

For rvs $X, Y$, it holds that $H(X,Y) = H(X) + H(Y|X)$.

- Proof immediately follow by the grouping axiom:

| $X \backslash Y$ | | | |
|---|---|---|---|
| | $P_{1,1}$ | $\ldots$ | $P_{1,n}$ |
| | $\vdots$ | $\vdots$ | $\vdots$ |
| | $P_{n,1}$ | $\ldots$ | $P_{n,n}$ |

Let $q_i = \sum_{j=1}^{n} p_{i,j}$ $\quad (= \Pr[X = i]$

$$H(P_{1,1}, \ldots, P_{n,n})$$
$$= H(q_1, \ldots, q_n) + \sum_i q_i H(\frac{P_{i,1}}{q_i}, \ldots, \frac{P_{i,n}}{q_i})$$
$$= H(X) + H(Y|X).$$

- Another proof. Let $(X,Y) \sim p$, and recall that $p(x,y) = p_X(x) \cdot p_{Y|X}(y|x)$.

$\implies \log p(x,y) = \log p_X(x) + \log p_{Y|X}(y|x)$

$\implies \operatorname{E} \log p(X,Y) = \operatorname{E} \log p_X(X) + \operatorname{E} \log p_{Y|X}(Y|X)$

$\implies H(X,Y) = H(X) + H(Y|X)$.

## $H(Y|X) \leq H(Y)$

Jensen inequality: for any concave function $f$, values $t_1, \ldots, t_k$ and $\lambda_1, \ldots, \lambda_k \in [0,1]$ with $\sum_i \lambda_i = 1$, it holds that $\sum_i \lambda_i f(t_i) \leq f(\sum_i \lambda_i t_i)$.
Let $(X, Y) \sim p$.

$$\begin{aligned}
H(Y|X) &= -\sum_{x,y} p(x,y) \log p_{Y|X}(y|x) \\
&= \sum_{x,y} p(x,y) \log \frac{p_X(x)}{p(x,y)} \\
&= \sum_{x,y} p_Y(y) \cdot \frac{p(x,y)}{p_Y(y)} \log \frac{p_X(x)}{p(x,y)} \\
&= \sum_{y} p_Y(y) \sum_{x} \frac{p(x,y)}{p_Y(y)} \log \frac{p_X(x)}{p(x,y)} \\
&\leq \sum_{y} p_Y(y) \log \sum_{x} \frac{p(x,y)}{p_Y(y)} \frac{p_X(x)}{p(x,y)} \\
&= \sum_{y} p_Y(y) \log \frac{1}{p_Y(y)} = H(Y).
\end{aligned}$$

## $H(Y|X) \leq H(Y)$ cont.

- Assume $X$ and $Y$ are independent (i.e., $p(x,y) = p_X(x) \cdot p_Y(y)$ for any $x, y$)

$\implies \quad p_{Y|X}(y|x) = p_Y(y)$ for any $x, y$

$\implies \quad H(Y|X) = H(Y)$

- Is the converse also true: $H(Y|X) = H(Y)$ implies $X$ and $Y$ are independent?

  Yes, since log is strictly concave in the range. Equality happens iff all $t_i$ are the same,

- which happens iff $p(x,y) = p_X(x)p_Y(y)$ for all $x, y$

# Other inequalities

- $H(X), H(Y) \leq H(X, Y) \leq H(X) + H(Y)$.
  Follows from $H(X, Y) = H(X) + H(Y|X)$.
  - Left inequality since $H(Y|X)$ is non negative.
  - Right inequality since $H(Y|X) \leq H(Y)$.
- $H(X, |Z) = H(X|Z) + H(Y|X, Z)$ (by chain rule)
- $H(X|Y, Z) \leq H(X|Y)$

Proof:

$$\begin{aligned}
H(X|Y,Z) &= \mathop{\mathrm{E}}_{(z,y)\leftarrow(Z,Y)} H(X|_{(Y,Z)=(z,y)}) \\
&= \mathop{\mathrm{E}}_{y\leftarrow Y} \mathop{\mathrm{E}}_{z\leftarrow Z|_{Y=y}} H(X|_{(Y,Z)=(z,y)}) \\
&= \mathop{\mathrm{E}}_{y\leftarrow Y} \mathop{\mathrm{E}}_{z\leftarrow Z|_{Y=y}} H((X|_{Y=y})|_{Z=z}) \\
&\leq \mathop{\mathrm{E}}_{y\leftarrow Y} H(X|_{Y=y}) \\
&= H(X|Y).
\end{aligned}$$

# Chain rule (for the entropy function), general case

**Claim 2**

For rvs $X_1, \ldots, X_k$, it holds that
$H(X_1, \ldots, X_k) = H(X_i) + H(X_2|X_1) + \ldots + H(X_k|X_1, \ldots, X_{k-1})$.

Proof: ?

- Extremely useful property!
- Analogously to the two variables case, it also holds that:
- $H(X_i) \leq H(X_1, \ldots, X_k) \leq \sum_i H(X_i)$
- $H(X_1, \ldots, X_K|Y) \leq \sum_i H(X_i|Y)$

## Examples

- (from last class) Let $X_1, \ldots, X_n$ be Boolean iid with $X_i \sim (\frac{1}{3}, \frac{2}{3})$. Compute $H(X_1, \ldots, X_n)$
- As above, but $X_n$ is set to $\bigoplus_{1 \le i \le n-1} X_i$ ?
  - Via chain rule?
  - Via mapping?

# Applications

- Let $X_1, \ldots, X_n$ be Boolean iids with $X_i \sim (p, 1-p)$ and let $X = X_1, \ldots, X_n$. Let $f$ be such that $\Pr[f(X) = z] = \Pr[f(X) = z']$, for every $k \in \mathbb{N}$ and $z, z' \in \{0,1\}^k$. Let $K = |f(X)|$.

  Prove that $\mathrm{E}\, K \leq n \cdot h(p)$.

- $$\begin{aligned} n \cdot h(p) &= H(X_1, \ldots, X_n) \\ &\geq H(f(X), K) \\ &= H(K) + H(f(X) \mid K) \\ &= H(K) + \mathrm{E}\, K \\ &\geq \mathrm{E}\, K \end{aligned}$$

- Interpretation

- Upper bounds

## Applications cont.

- How many comparisons it takes to sort $n$ elements?

  Let S be a sorter for $n$ elements algorithm making $t$ comparisons.

  What can we say about $t$?
- Let $X$ be a uniform random permutation of $[n]$ and let $Y_1, \ldots, Y_t$ be the answers S gets when sorting $X$.
- $X$ is determined by $Y_1, \ldots, Y_t$.

  Namely, $X = f(Y_1, \ldots, Y_t)$ for some function $f$.
- $H(X) = \log n!$
- $$\begin{aligned} H(X) &= H(f(Y_1, \ldots, Y_t)) \\ &\leq H(Y_1, \ldots, Y_t) \\ &\leq \sum_i H(Y_i) \\ &\leq t \end{aligned}$$

$\implies \quad t \geq \log n! = \Theta(n \log n)$

## Concavity of entropy function

Let $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_n)$ be two distributions, and for $\lambda \in [0,1]$ consider the distribution $\tau_\lambda = \lambda p + (1-\lambda)q$.
(i.e., $\tau_\lambda = (\lambda p_1 + (1-\lambda)q_1, \ldots, \lambda p_n + (1-\lambda)q_n)$.

**Claim 3**

$H(\tau_\lambda) \geq \lambda H(p) + (1-\lambda)H(q)$

Proof:

- Let $Y$ over $\{0,1\}$ be $0$ wp $\lambda$
- Let $X$ be distributed according to $p$ if $Y = 0$ and according to $q$ otherwise.
- $H(\tau_\lambda) = H(X) \geq H(X \mid Y) = \lambda H(p) + (1-\lambda)H(q)$

We are now certain that we drew the graph of the (two-dimensional) entropy function right...

Part II

# Mutual Information

# Mutual information

- $I(X; Y)$ — the "information" that $X$ gives on $Y$
- $$\begin{aligned} I(X; Y) &:= H(Y) - H(Y|X) \\ &= H(Y) - (H(X, Y) - H(X)) \\ &= H(X) + H(Y) - H(X, Y) \\ &= H(X) - H(X|Y) \\ &= I(Y; X). \end{aligned}$$
- The mutual information that $X$ gives about $Y$ equals the mutual information that $Y$ gives about $X$.
- $I(X; Y) \geq 0$. When $0$?
- $I(X; X) = H(X)$
- $I(X; f(X)) = H(f(X))$ (and smaller then $H(X)$ if $f$ is non-injective)
- $I(X; Y, Z) \geq I(X; Y), I(X; Z)$ (since $H(X \mid Y, Z) \leq H(X \mid Y), H(X \mid Z)$)
- $I(X; Y|Z) := H(Y|Z) - H(Y|X, Z)$ $\geq 0$
- $I(X; Y|Z) = I(Y; X|Z)$ (since $I(X'; Y') = I(Y'; X')$)

# Numerical example

- Example

| $X \backslash Y$ | 0 | 1 |
|---|---|---|
| 0 | $\frac{1}{4}$ | $\frac{1}{4}$ |
| 1 | $\frac{1}{2}$ | 0 |

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= 1 - \frac{3}{4} \cdot h(\frac{1}{3}) \\ &= I(Y;X) \\ &= H(Y) - H(Y|X) \\ &= h(\frac{1}{4}) - \frac{1}{2} h(\frac{1}{2}) \end{aligned}$$

# Chain rule for mutual information

**Claim 4 (Chain rule for mutual information)**

For rvs $X_1, \ldots, X_k, Y$, it holds that
$I(X_1, \ldots, X_k; Y) = I(X_1; Y) + I(X_2; Y|X_1) + \ldots + I(X_k; Y|X_1, \ldots, X_{k-1})$.

Proof: ? HW

## Examples

- Let $X_1, \ldots, X_{n-1}$ be iid uniform bits (i.e., $X_i \sim (\frac{1}{2}, \frac{1}{2})$), and let $X_n = \bigoplus_{i \in [n-1]} X_i$. Compute $I(X_1, \ldots, X_{n-1}; X_n)$.
  - Directly,
    $I(X_1, \ldots, X_{n-1}; X_n) = H(X_n) - I(X_n | X_1, \ldots, X_{n-1}) = 1 - 0 = 1$
  - Using chain rule,

$$
\begin{aligned}
& I(X_1, \ldots, X_{n-1}; X_n) \\
& = I(X_1; X_n) + I(X_2; X_n | X_1) + \ldots + I(X_{n-1}; X_n | X_1, \ldots, X_{n-2}) \\
& = 0 + 0 + \ldots + 1 = 1.
\end{aligned}
$$

- Let $T$ and $F$ be the top and front side, respectively, of a $6$-sided fair dice. Compute $I(T; F)$.

$$
\begin{aligned}
I(T; F) &= H(T) - H(T|F) \\
&= \log 6 - \log 4 \\
&= \log 3 - 1.
\end{aligned}
$$

Part III

# Data processing

## Data processing Inequality

**Definition 5 (Markov Chain)**

Rvs $(X, Y, Z) \sim p$ form a Markov chain, denoted $X \to Y \to Z$, if $p(x, y, z) = p_X(x) \cdot p_{Y|X}(y|x) \cdot p_{Z|Y}(z|y)$, for all $x, y, z$.

Example: random walk on graph.

**Claim 6**

If $X \to Y \to Z$, then $I(X; Y) \geq I(X; Z)$.

- By Chain rule, $I(X; Y, Z) = I(X; Z) + I(X; Y|Z) = I(X; Y) + I(X; Z|Y)$.1
- $I(X; Z|Y) = 0$
  - $p_{Z|_{Y=y}} \equiv p_{Z|_{Y=y,X=x}}$ for any $x, y$
  - $$\begin{aligned} I(X; Z|Y) &= H(Z|Y) - H(Z|Y, X) \\ &= \mathop{\mathrm{E}}_{y \leftarrow Y} H(p_{Z|_{Y=y}}) - \mathop{\mathrm{E}}_{(x,y) \leftarrow (Y,X)} H(p_{Z|_{Y=y,X=x}}) \\ &= \mathop{\mathrm{E}}_{y \leftarrow Y} H(p_{Z|_{Y=y}}) - \mathop{\mathrm{E}}_{y \leftarrow Y} H(p_{Z|_{Y=y}}) = 0. \end{aligned}$$
- Since $I(X; Y|Z) \geq 0$, we conclude $I(X; Y) \geq I(X; Z)$.$\square$

# Fano's Inequality

- How well can we guess $X$ from $Y$?
- Could with no error if $H(X|Y) = 0$. What if $H(X|Y)$ is small?

**Theorem 7 (Fano's inequality)**

*For any rvs $X$ and $Y$, and any (even random) $g$, it holds that*

$$h(P_e) + P_e \log|\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y)$$

*for $\hat{X} = g(Y)$ and $P_e = \Pr\left[\hat{X} \neq X\right]$.*

- Note that $P_e = 0$ implies that $H(X|Y) = 0$
- The inequality can be weekend to $1 + P_e \log|\mathcal{X}| \geq H(X|Y)$,
- Alternatively, to $P_e \geq \frac{H(X|Y)-1}{\log|\mathcal{X}|}$
- Intuition for $\propto \frac{1}{\log|\mathcal{X}|}$
- We call $\hat{X}$ an estimator for $X$ (from $Y$).

## Proving Fano's inequality

Let $X$ and $Y$ be rvs, let $\hat{X} = g(Y)$ and $P_e = \Pr\left[\hat{X} \neq X\right]$.

- Let $D = \begin{cases} 1, & \hat{X} \neq X \\ 0, & \hat{X} = X. \end{cases}$

$$\begin{aligned} H(D, X|\hat{X}) &= H(X|\hat{X}) + \underbrace{H(D|X, \hat{X})}_{=0} \\ &= \underbrace{H(D|\hat{X})}_{\leq H(D) = h(P_e)} + \underbrace{H(X|D, \hat{X})}_{\leq P_e \log|\mathcal{X}|(?)} \end{aligned}$$

- It follows that $h(P_e) + P_e \log|\mathcal{X}| \geq H(X|\hat{X})$
- Since $X \to Y \to \hat{X}$, it holds that $I(X; Y) \geq I(X; \hat{X})$
  $\implies H(X|\hat{X}) \geq H(X|Y)$